Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using ^{1}H-NMR-based metabonomics

JOANNE T. BRINDLE[1], HENRIK ANTTI[1], ELAINE HOLMES[1], GEORGE TRANTER[1], JEREMY K. NICHOLSON[1], HUGH W. L. BETHELL[2], SARAH CLARKE[2], PETER M. SCHOFIELD[2], ELAINE MCKILLIGIN[3], DAVID E. MOSEDALE[4] & DAVID J. GRAINGER[4]

[1]Biological Chemistry, Biomedical Sciences Division, Faculty of Medicine, Imperial College of Science, Technology and Medicine, Sir Alexander Fleming Building, Exhibition Road, South Kensington, London, UK

[2]Department of Cardiology, Papworth Hospital NHS Trust, Cambridge, UK

[3]GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Stevenage, UK

[4]Department of Medicine, Box 157, Addenbrooke's Hospital, Cambridge, UK

Correspondence should be addressed to D.J.G.; e-mail: djg15@cam.ac.uk

Published online: 25 November 2002, doi:10.1038/nm802

Although a wide range of risk factors for coronary heart disease have been identified from population studies, these measures, singly or in combination, are insufficiently powerful to provide a reliable, noninvasive diagnosis of the presence of coronary heart disease. Here we show that pattern-recognition techniques applied to proton nuclear magnetic resonance (^{1}H-NMR) spectra of human serum can correctly diagnose not only the presence, but also the severity, of coronary heart disease. Application of supervised partial least squares-discriminant analysis to orthogonal signal-corrected data sets allows >90% of subjects with stenosis of all three major coronary vessels to be distinguished from subjects with angiographically normal coronary arteries, with a specificity of >90%. Our studies show for the first time a technique capable of providing an accurate, noninvasive and rapid diagnosis of coronary heart disease that can be used clinically, either in population screening or to allow effective targeting of treatments such as statins.

Coronary heart disease (CHD) is a major cause of mortality and morbidity in developed countries, affecting as many as one in three individuals before the age of 70 years[1]. Over the past three decades a range of environmental and biochemical risk factors for the development of CHD have been identified in cross-sectional studies[2]. For example, tobacco smoking is associated with an approximately two-fold increased risk of CHD[3]. Similarly, high levels of cholesterol in large, triglyceride-rich lipoprotein particles (mainly very low-density lipoprotein (VLDL) and low-density lipoprotein (LDL)) and lower levels of cholesterol in high-density lipoprotein (HDL) particles are known to be associated with increased risk of CHD[4].

These epidemiological studies have been very useful in several ways. First, they have underpinned public health policy on a range of issues, discouraging tobacco smoking and promoting a low-cholesterol diet[5]. Second, they have provided vital clues as to the underlying molecular mechanisms that cause atherosclerosis and CHD[6]. However, the risk factors identified so far from cross-sectional epidemiological studies are insufficiently powerful to provide a clinically useful diagnosis of CHD. Although algorithms have been designed based on a range of risk factors, such as age, sex, lipoprotein levels and blood pressure, which can identify subpopulations at very significant excess risk of CHD, even the best of these based on the excellent Prospective Cardiovascular Münster (PROCAM) study in Münster, Germany, cannot diagnose the presence of CHD on an individual-by-individual basis[7,8].

Recently, however, there have been technical advances that have allowed extremely high-density data sets to be constructed from individuals. Techniques such as genomics, proteomics and metabonomics (a systems approach to examining the changes in hundreds or thousands of low-molecular-weight metabolites in an intact tissue or biofluid[9]) offer the prospect of efficiently distinguishing individuals with particular disease or toxic states. Of these techniques, NMR-based metabonomics offers several distinct advantages in a clinical setting. First, it can be carried out on standard preparations of serum, plasma or urine[10,11], circumventing the need for specialist preparations of cellular RNA and protein required for genomics and proteomics, respectively[12-14]. Second, many of the risk factors already identified (such as levels of various lipids) are small-molecule metabolites that will contribute to the metabonomic data set.

In this study we have applied recently developed pattern-recognition techniques to NMR spectra of either serum or plasma taken from individuals who have been extensively characterized, both for the presence of CHD by the gold-standard angiographic technique and for a wide range of conventional risk factors. This allows direct comparison of the performance of the metabonomic analysis as a diagnostic technique with algorithms based on conventional risk factors.

The 600-MHz ^{1}H-NMR spectra of human sera from patients with severe CHD (triple vessel disease (TVD) patients; $n = 36$) and patients with angiographically normal coronary arteries (NCA patients; $n = 30$) were compared visually (Figs. 1*a* and *b*). The clinical characteristics of the populations studied here are provided in the Supplemental Note online. Few systematic differences were detected when the two groups were compared visually. Chemical components were assigned to the spectra on the basis of previously published

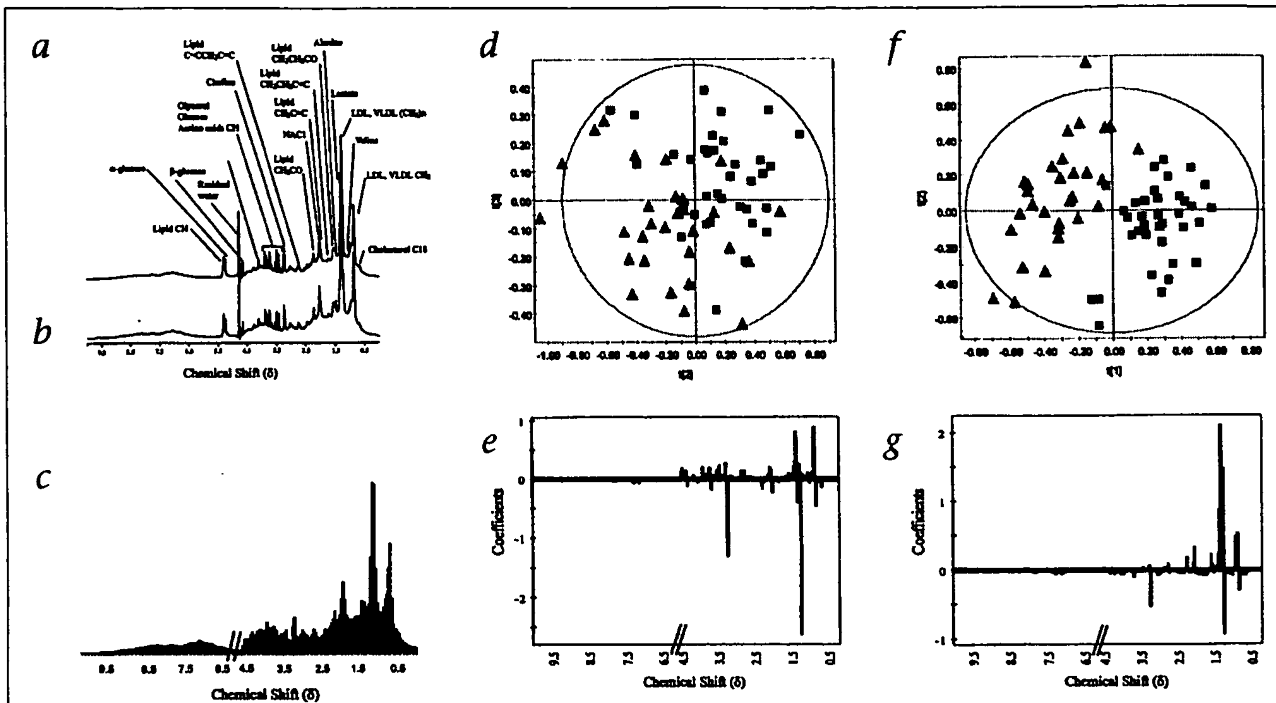

Fig. 1 Comparison of patients with severe atherosclerosis (TVD) and patients with normal coronary arteries (NCA). The 600 MHz ^{1}H-NMR spectra of serum samples from a typical NCA patient (*a*) and a TVD patient (*b*) are shown. The chemical shifts of a selection of major metabolites are indicated (based on assignments from 2D-TOCSY spectra), although these metabolites do not contribute all (or in some cases, even most) of the signal at the indicated chemical shift. ***c***, Data-reduced ^{1}H-NMR spectrum of a serum sample from a typical TVD patient. The region between δ4.5 and δ6 has been deleted to reduce the likelihood of any variance contribution from incomplete suppression of the water signal. ***d***, PLS-DA scores plot showing the considerable separation achieved between NCA (▲) and TVD (■) samples. Note that optimum separation occurred in the second and third principal components (t[2] and t[3]). ***e***, The regression coefficients of the PLS-DA model shown in (*d*). Positive coefficients indicate relatively higher values for that spectral region in the TVD samples compared with the NCA samples, whereas negative coefficients indicate lower values. The magnitude of the coefficient represents the relative importance of each data bin on the separation achieved in (*d*). ***f***, PLS-DA scores plot after application of the OSC data filter to remove uncorrelated variance components. Note the considerable improvement in separation achieved (compared with *d* above), which now occurs in the first two principal components (t[1] and t[2]). ***g***, The regression coefficients for the PLS-DA model using the OSC transformed data set.

data[15,16]. To reduce the complexity of the NMR data to facilitate pattern recognition, the spectra were automatically data-reduced to 245 integral segments, each comprising 0.04 p.p.m., before chemometric analysis (Fig. 1*c*).

To determine whether it was possible to distinguish TVD and NCA patients on the basis of the NMR spectra, we carried out principal components analysis[17] (PCA) and partial least squares-discriminant analysis[17] (PLS-DA). The PLS-DA scores plot of the second and third principal components (PC2 and PC3) shows that, although there was overlap between the two sample classes, some clustering was evident (Fig. 1*d*). The regions of the NMR spectrum that most strongly influence separation between NCA and TVD samples are indicated by the regression coefficients (Fig. 1*e*). The coefficients were derived from the PLS-DA model and each bar represents a spectral region covering 0.04 p.p.m., showing how the ^{1}H-NMR profile of the TVD samples differed from the ^{1}H-NMR profile of the NCA serum samples. A positive value indicated there was a relatively greater concentration of metabolite (assigned using NMR chemical shift assignment tables) present in TVD samples and a negative value indicated a relatively lower concentration, with respect to NCA samples. In general, the regression coefficients, or loadings, most influential for the TVD samples lie around δ0.86 (due mainly to CH_3 groups from fatty acid side chains in lipids, in particular LDL and VLDL) and δ1.26, δ1.3 and δ1.34 (due mainly to $(CH_2)_n$ groups from fatty acid side chains in lipids, in particular VLDL and LDL). The loadings most influential for the NCA samples lie around δ1.22 (due mainly to $(CH_2)_n$ groups from fatty acid side chains in lipids, in particular HDL) and δ3.22 (due to choline $-N(CH_3)_3^+$). The region at δ3.22 is assigned to $-N(CH_3)_3^+$ groups in molecules containing the choline moiety, principally phosphatidylcholine from lipoproteins, mainly HDL, based on the known phospholipid content of lipoproteins.

If chemometric analysis is suggestive of separation between the classes under investigation, orthogonal signal correction (OSC) can be used to optimize the separation[18], thus improving the performance of subsequent multivariate pattern-recognition analysis and enhancing the predictive power of the model. After application of OSC, the TVD and NCA groups were well separated in the PLS-DA scores plot of PC1 and PC2 (Fig. 1*f*). The regression coefficients (Fig. 1*g*) indicated that the same regions of the spectra that contributed to the clustering in the unfiltered data set also contributed to the clustering seen after application of OSC. The statistically significant loadings are presented in the Supplemental Note online.

Approximately 80% of the samples (the 'training set') were then selected at random to construct a PLS-DA model that could then be used to predict the class membership of the remaining 20% of samples (the 'test set'). The regression coefficients for the training set again indicated that the same spectral regions contributed most strongly to the discrimination of the classes: lipids, mostly VLDL, LDL and HDL, and choline. The PLS-DA model calculated from OSC-filtered ^{1}H-NMR data for the training set predicted the presence of CHD with a sensitivity of 92% and a specificity of 93% based on a 99% confidence limit for class membership (Fig. 2). The Y-predicted scatter plot assigned samples to either class 1 (TVD) or class 0 (NCA) using an *a priori* cut-off of 0.5, and showed the ability of ^{1}H-NMR-based metabonomics to predict class membership (NCA or TVD) of unknown samples.

As the cohort of patients under investigation was selected only on the basis of coronary artery disease status, the proportion of men and women in the two groups and their age distributions are different (see Supplemental Note online). Such gender bias is inevitable if we are to avoid overemphasizing the power of our diagnostic assay by artifactually removing major sources of variation, as commonly occurs when using highly age- and sex-matched groups. However, the classification of the patients as NCA or TVD on the basis of the NMR spectrum did not depend on the sex bias of the patient groups: most male NCA subjects were correctly classified as NCA, despite the fact that the NCA group was primarily composed of women (Fisher's exact test; $P = 0.006$). In contrast, if gender rather than artery status is used to categorize the samples, PLS-DA is able to classify the individuals on the basis of sex with 100% sensitivity and specificity (data not shown). This demonstrates that our metabonomic analysis is able to diagnose the presence of CHD against the background variation in the gender distribution of the groups, but is not relying on gender differences to make the diagnosis.

This study demonstrated that ^{1}H-NMR-based metabonomic analysis of serum samples, in itself minimally invasive and non-destructive of sample, can achieve a clinically useful diagnostic performance, when compared with invasive angiography.

To determine whether ^{1}H-NMR-based metabonomic analysis could distinguish the severity of CHD present, we collected a set of samples from individuals with stenosis of one (mild, $n = 28$), two (moderate, $n = 20$) or three (severe, $n = 28$) major coronary arteries. Although this is only a crude indicator of disease severity, it is plausible that the number of stenosed vessels correlated (at least weakly) with atherosclerotic plaque load in the whole body. The 600-MHz ^{1}H-NMR spectra from the 76 patients with CHD of varying severity were obtained and analyzed by PCA and PLS-DA. Following application of OSC to ^{1}H-NMR data, separation between the mild, moderate and severe CHD samples was evident and the regression coefficients indicated that once again the lipid parameters contributed most strongly to the separation (Figs. 3*a* and *b*). To optimally view the separation between the mild, moderate and severe CHD samples, PLS-DA models and regression coefficients were calculated for mild and moderate CHD (Figs. 3*c* and *d*), moderate and severe CHD (Figs. 3*e* and *f*), and mild and severe CHD (Figs. 3*g* and *h*). For the models shown in Figs. 3*a*, *c* and *g*, the second principal component was not statistically significant in PLS-DA, but is shown for clarity.

The mild, moderate and severe CHD samples were also compared on the basis of established clinical risk factors (tabulated in full in the Supplemental Note online). None of the risk factors measured (including age, blood pressure, LDL and HDL cholesterol, total cholesterol, total triglyceride, fibrinogen, plasminogen activator inhibitor (PAI-1), white blood cell count, creatinine or history of cigarette smoking) was significantly different between the three groups ($P > 0.05$ by ANOVA in each case; Supplemental Note online). Furthermore, chemometric analysis of these clinical data was not able to determine the number of stenosed vessels. In the PLS-DA model of the clinical data, none of the principal components extracted was statistically significant, in contrast to the PLS-DA model based on the NMR data shown in Fig. 3. We conclude that ^{1}H-NMR-based metabonomics is better able to distinguish the severity of CHD based on a single blood sample than any of the conventional risk factors yet identified, even when pattern-recognition methodology is applied.

Discussion

We have demonstrated that it is possible to completely separate CHD patients with stenosis of all three major arteries from subjects with normal coronary arteries using both unsupervised PCA and supervised PLS-DA applied to ^{1}H-NMR spectra of human serum. Furthermore, using the supervised PLS-DA algorithm, it is possible to predict the artery status of unknown samples using a training set composed of only 24 individuals with NCA and 30 individuals with TVD. The small size of the training set required to achieve greater than 90% sensitivity and specificity highlights the power of this technique. Substantially larger training sets obtained through application of this technique to clinical practice should further improve the diagnostic sensitivity and specificity of the

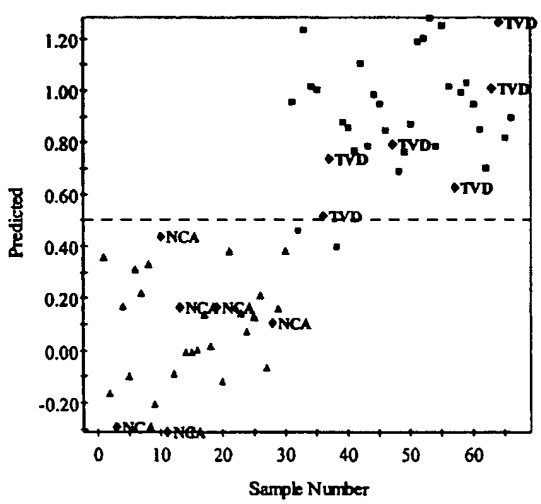

Fig. 2 Prediction of coronary artery status using the PLS-DA model. A PLS-DA model was constructed using the OSC filtered data from 24 NCA patients (▲) and 30 TVD patients (■) (the 'training set'). This model was then used to 'predict' the coronary artery status of a further six samples of each class that were not used in the construction of the model (the 'test set'). Predictions are made using a Y-predicted scatter plot with the *a priori* cut-off of 0.5 for class membership. The test sets are shown as black diamonds, with their angiographically determined artery status shown in text.

technique. Both PCA and PLS-DA analyses were improved by prior application of the OSC technique to the data set[18]. On this basis, OSC is likely to find widespread application in pattern recognition for high-information-density data sets, not only in metabonomics, but also in genomics and proteomics.

Both of the pattern-recognition algorithms used here rely on extraction of linear associations between the input variables, which can significantly limit the power of the analysis (the strengths and weaknesses of our study design are discussed in detail in the Supplemental Note online). It is already clear that neural-network–based pattern-recognition techniques can considerably improve the ability to classify individuals on the basis of many interrelated input variables[19], particularly when membership in a class (such as having CHD) may result from one of a range of unrelated causes. Nevertheless, the methods we applied are sufficiently powerful to allow classification of the individuals we studied, and provide one additional benefit over neural network methods: they allow information to be more easily gained as to what aspects of the input data set were particularly important in allowing the classification to be made. These regression coefficients (Figs. 1*e* and *g*) indicated an important contribution from two of the data regions in particular: the bins around chemical shifts of δ1.30 and δ3.22. Although the peaks around δ1.30 are known to result from lipid CH_2 resonances and to correlate with the levels of LDL-cholesterol ($r = 0.45$; $P < 0.01$), it is notable that only 20–30% of the variance in this bin is related to classical measurement of LDL-cholesterol concentration. The remaining variance is likely to result from subtle chemical differences in the lipid composition of LDL particles between individuals, for example, the degree of fatty-acid side-chain unsaturation and lipoprotein–protein molecular interactions. The assignment of a metabolite to a particular chemical shift only indicates that the concentration of that metabolite contributes to the variance seen at that chemical shift. However, that contribution may only be a small part of the total variance, particularly at low chemical shifts where many different molecular species contribute to each spectral interval. The importance of bins with a variance component due to lipoprotein composition will likely contribute to ongoing studies using both NMR and other analytical techniques to understand the contribution of lipoprotein particle composition to the development of CHD[20]. It does, however, em-

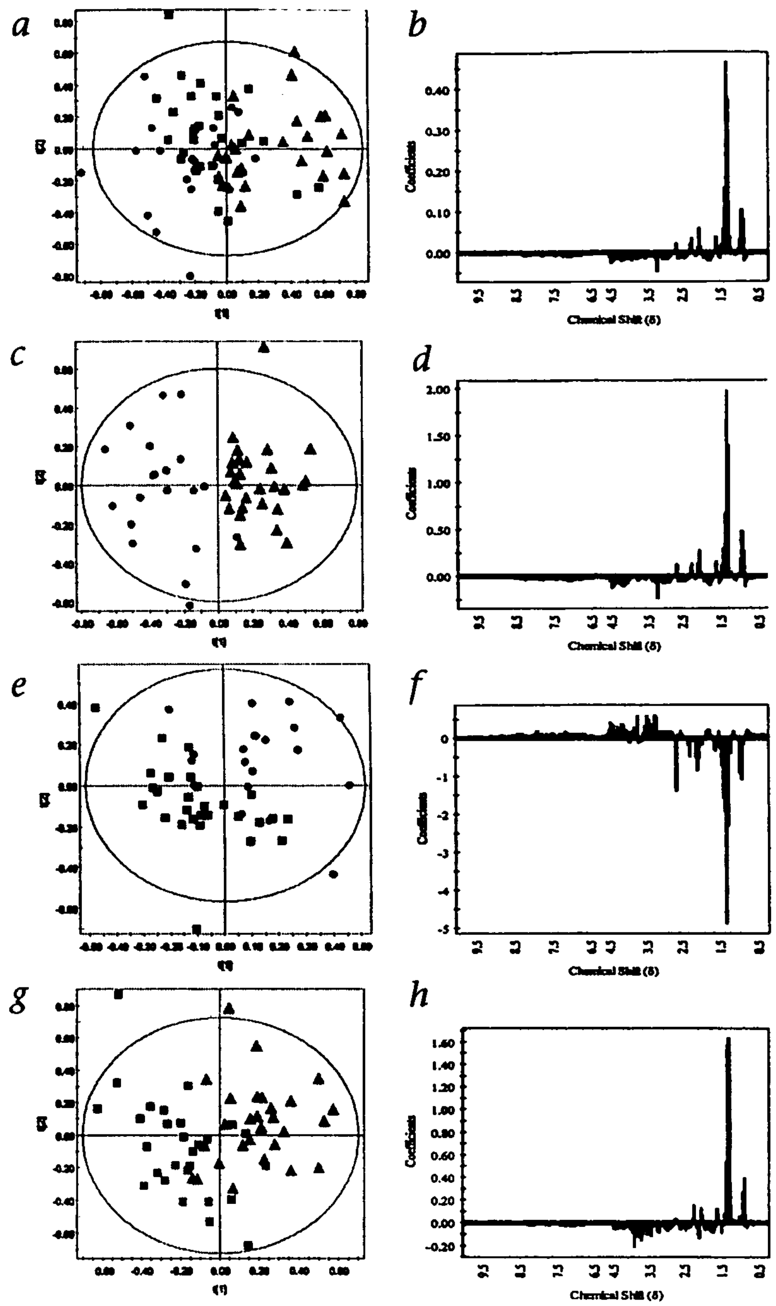

Fig. 3 Comparison of patients with different severity of coronary atherosclerosis. Spectra were generated for a further 76 males all of whom had angiographically proven coronary artery disease. These individuals were classified as having mild (n = 28), moderate (n = 20) or severe (n = 28) disease according to whether they had stenosis (>50% blockage) of one, two or three of the major coronary arteries. PLS-DA models were then generated on the OSC-transformed dataset exactly as for Fig. 1. (***a***) PLS-DA model comparing all three severity groups, mild (▲), moderate (●) or severe (■). Individuals with mild disease are well separated from the more severe groups in the first principal component. (***b***) Regression coefficient plot for the PLS-DA model shown in (*a*). (***c*** and ***d***) PLS-DA model scores plot and regression coefficient plot, respectively, produced using only the mild and moderate groups. (***e*** and ***f***) PLS-DA model scores plot and regression coefficient plot, respectively, produced using only the moderate and severe groups. (***g*** and ***h***) PLS-DA model scores plot and regression coefficient plot, respectively, produced using only the mild and severe groups. In each case, positive regression coefficients indicate higher signal in the more severe group. Note that the second principal component (t[2]) is not statistically significant for the models shown in (*a*), (*c*) and (*g*).

Methods

Sample collection and preparation. Patients in the TVD group had significant coronary artery disease (defined as a reduction of more than 50% in the intralumenal diameter) of all three major coronary arteries. Patients in the NCA group had chest pain and a positive exercise test, but normal coronary arteries by angiography. The inclusion and exclusion criteria we applied are described in detail in the Supplemental Note online. Consecutive patients presenting at Papworth Hospital (Cambridgeshire, UK) who met the criteria for either the TVD or NCA group were recruited to the study. The clinical data for these patient groups are provided in the Supplemental Note online.

Blood was drawn from each patient and allowed to clot in plastic tubes for 2 h at room temperature, and the serum was collected by centrifugation. Aliquots of serum were stored at −80 °C until assayed.

For the study of extent of CHD, patients were recruited according to the same criteria, except that patients with more than 50% stenosis of one, two or all three coronary arteries (assessed by two independent observers) were recruited and females were excluded. Blood samples from these patients were drawn into Diatube H tubes, and platelet-poor plasma was prepared as described above. Aliquots of plasma were stored at −80 °C until assayed.

NMR spectroscopy. Before NMR analysis, samples (150 μl) were diluted with solvent solution (10% ^2H_2O v/v; 0.9% NaCl w/v (350 μl)) in 5-mm high-quality NMR tubes (Goss Scientific Instruments, Great Baddow, Essex, UK). Conventional ^{1}H-NMR spectra of the blood samples were measured at 600.22 MHz on a Bruker Avance-600 spectrometer (Billerica, Massachusetts) operating at 600 MHz ^{1}H frequency, using the pulse sequence: RD - 90° - t_1 - 90° - t_m - 90° - acquire free induction decay (FID). RD represents a relaxation delay of 1.5 s during which the water resonance is selectively irradiated, and t_1 corresponds to a fixed interval of 4 μs. The water resonance is irradiated for a second time during the mixing time (t_m, 150 ms). For each sample, 64 FIDs were collected into 32,768 data points using a spectral width of 8389.3 Hz and an acquisition time of 1.95 s. The FIDs were multiplied by an exponential weighting function corresponding to a line broadening of 0.3 Hz before Fourier transformation. The acquired NMR spectra were corrected for phase and baseline distortions using XWINNMR (version 2.1, Bruker) and referenced to lactate (CH_3 δ1.33).

Data reduction of NMR data. The ^{1}H-NMR spectra (δ10–0.2) were automatically data-reduced to 245 integral segments of equal length (δ0.04) using AMIX (Analysis of MIXtures software package, version 2.5, Bruker). Each segment consisted of the integral of the NMR region to which it was associated. To remove the effects of variation in the suppression of the water resonance, and also the effects of variation in the urea signal caused by partial cross-solvent saturation by means of solvent-exchanging protons, the region δ6.0 to 4.5 was set to zero integral. The data were normalized to total spectral area and centered scaling was applied.

OSC of NMR data. PCA and PLS-DA were carried out on the ^{1}H-NMR data both before and after OSC. The class identity is used as a response vector, Y, to describe the variation between the sample classes. The OSC method[18] then locates the longest vector describing the variation between the samples that is not correlated with the Y-vector and removes it from the data matrix. The resultant data set has been filtered to allow pattern recognition focused on the variation correlated to features of interest within the sample population, rather than non-correlated, orthogonal variation.

PCA. PCA is a bilinear decomposition method used for overviewing 'clusters' within multivariate data[17]. The data (X) are represented in K-dimensional space (where K is equal to the number of variables) and reduced to a few principal components (or latent variables) that describe the maximum variation within the data, independent of any knowledge of class membership (that is, 'unsupervised'). The principal components are displayed as a set of 'scores' (t), which highlight clustering or outliers, and a set of 'loadings' (p), which highlight the influence of input variables on t.

PLS-DA. PLS-DA is a multivariate classification method based on PLS, the regression extension of PCA[17]. Whereas PCA works to explain maximum variation between samples, PLS-DA explains maximum separation between defined class samples in the data (X); for this reason, class membership of each sample must be known. PLS-DA is done by a PLS regression against a 'dummy matrix' (Y), which describes variation according to class. Variation is interpreted in terms of X- and Y-scores (T, U), X-loadings (P), X- and Y-weights (W, C), and PLS regression coefficients (B). Once a PLS-DA model is calculated and validated it can be used for prediction of class membership for unknown samples.

phasize an important facet of high-data-density metabolic analysis in that it is entirely unnecessary to understand fully the complex molecular differences that underlie the spectral features associated with CHD to be able to correctly classify individuals with very high sensitivity and specificity. Further analysis of the molecular basis of the spectral differences, however, will give insight into the mechanistic processes involved.

Whereas currently a firm diagnosis of CHD can only be made through application of angiography, which is both expensive and invasive, the introduction of metabonomic screening would allow diagnosis to be made simply and cheaply on the basis of a single blood sample. As angiographic status was used as the classification variable during model construction, the diagnostic assay reported here behaves as a simple replacement for angiography. Like angiography, it does not distinguish stable and unstable angina, which may be a major factor in determining the likelihood of future myocardial infarction. Nevertheless, the availability of a relatively cheap and noninvasive replacement for angiography would revolutionize the provision of health care for CHD, allowing both widespread population screening and more efficient targeting of drugs, such as statins. These drugs, although broadly effective in reducing the risk of myocardial infarction, are difficult to target to those most in need of treatment.

Competing interest statement

The authors declare competing financial interests: see the website (http://www.nature.com/naturemedicine) for details.

1. National Statistics Series DH1 no. 31, Mortality Statistics (1998). The Stationery Office (UK Government), London, England.
2. Kjelsberg, M.O., Cutler, J.A. & Dolecek, T.A. Brief description of the Multiple Risk Factor Intervention Trial. *Am. J. Clin. Nutr.* 65 (Suppl. 1), 191S–195S (1997).
3. Kuller, L.H. *et al.* Cigarette smoking and mortality. MRFIT Research Group. *Prev. Med.* 20, 638–654 (1991).
4. Multiple Risk Factor Intervention Trial Research Group. Relationship between baseline risk factors and coronary heart disease and total mortality in the Multiple Risk Factor Intervention Trial. *Prev. Med.* 15, 254–273 (1986).
5. McIlvain, H.E., McKinney, M.E., Thompson, A.V. & Todd, G.L. Application of the MRFIT smoking cessation program to a healthy, mixed-sex sample. *Am. J. Prev. Med.* 8, 165–170 (1992).
6. Ross, R. Atherosclerosis—an inflammatory disease. *N. Engl. J. Med.* 340, 115–126 (1999).
7. Cullen, P., Funke, H., Schulte, H. & Assmann, G. Lipoproteins and cardiovascular risk—from genetics to CHD prevention. *Eur. Heart J.* 19 (Suppl. C), C5–C11 (1998).
8. Isles, C.G. & Paterson, J.R. Identifying patients at risk for coronary heart disease: implications from trials of lipid-lowering drug therapy. *Q.J. Med.* 93, 567–574 (2000).
9. Nicholson, J.K., Lindon, J.C. & Holmes, E. 'Metabonomics': Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29, 1181–1189 (1999).
10. Nicholson, J.K. & Wilson, I.D. High resolution proton magnetic resonance spectroscopy of biological fluids *Prog. Nucl. Magn. Reson. Spectrosc.* 21, 449–501 (1989).
11. Lindon, J.C., Holmes, E. & Nicholson, J.K. Pattern recognition methods and applications in biomedical magnetic resonance. *Prog. Nucl. Magn. Reson. Spectrosc.* 39, 1–40 (2001).
12. Lindon, J.C., Nicholson, J.K.N., Holmes, E. & Everett, J.R. Metabonomics: metabolic processes studied by NMR spectroscopy of biofluids. *Concepts Magn. Reson.* 12, 289–320 (2000).
13. Holmes, E. *et al.* Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chem. Res. Toxicol.* 13, 471–478 (1999).
14. Storck, T., von Brevern, M.C., Behrens, C.K., Scheel, J. & Bach, A. Transcriptomics in predictive toxicology. *Curr. Opin. Drug Discov. Devel.* 5, 90–97 (2002).
15. Nicholson, J.K., Foxall, P.J., Spraul, M., Farrant, D.R. & Lindon, J.C. 750 MHz ^{1}H and ^{1}H-^{13}C NMR spectroscopy of human blood plasma. *Anal. Chem.* 67, 793–811 (1995).
16. Ala-Korpela, M. ^{1}H-NMR spectroscopy of human blood plasma. *Prog. Nucl. Magn. Reson. Spectrosc.* 27, 475–554 (1995).
17. Eriksson, L., Johansson, E., Kettanah-Wold, N. & Wold, S. *Introduction to Multi and Megavariate Data Analysis Using Projection Methods (PCA and PLS-DA)* (Umetrics AB, Malmo, Sweden, 1999).
18. Wold, S., Antti, H., Lindgren, F. & Ohman, J. Orthogonal signal correction of near-infrared spectra. *Chemometrics Intelligent Lab. Systems* 44, 175–185 (1998).
19. Ala-Korpela, M., Hiltunen, Y. & Bell, J.D. Quantification of biomedical NMR data using artificial neural network analysis: Lipoprotein lipid profiles from 1H NMR data of human plasma. *NMR Biomed.* 8, 235–244 (1995).
20. Otvos, J. In *Handbook of Lipoprotein Testing* (eds. Rifai, N., Warnick, R. & Dominiczak, M.) 497–508 (AACC Press, Washington DC, 1997).